Natural Language Processing Final Project Report

By ACK-Comets

Table of Contents

1. Overview

This section deals with a detailed description of the problem statement as well an overview of the proposed solution architecture.

1.1. Problem Description

The objective of this project was to create a sophisticated search engine that would yield high-quality search results by combining NLP features in addition to primitive keyword search. In the subsequent sections, the report will compare and contrast the two strategies through a few sample queries to highlight differences in search quality.

1.2. Proposed Solution

The solution process consisted of the following tasks:

Task 1

The goal of this task was to collect a corpus of news articles that met the following conditions:

- Minimum of 1000 articles
- Minimum of 100,000 words

Task 2

The goal of this task was to create a simple search pipeline that only used keywords. It was divided into the following subtasks:

- *Search Index Creation*
  - Segmented news articles into sentences
  - Tokenized sentences into words
  - Indexed word vector per sentence into search engine

- *Natural Language Query Parsing and Search*
  - Segmented user query into sentences
  - Tokenized sentences into words
  - Ran a search query using the query keyword vector against the index

- *Evaluation*
  - Evaluated the top-10 results of 10 sample queries

Task 3

The goal of this task was to create a richer pipeline that used NLP features. It was divided into the following subtasks:

- *Search Index Creation*
  - Segmented news articles into sentences
  - Tokenized sentences into words
  - Extracted following NLP features:
    - Lemmas
    - Stems
    - Part-Of-Speech Tags
    - Head Word
    - Hypernyms
    - Hyponyms
    - Meronyms
    - Holonyms
  - Indexed feature vector per sentence into search engine

- *Natural Language Query Parsing and Search*
  - Segmented user query into sentences
  - Tokenized sentences into words
  - Ran a search query using the query keyword vector against the index

- *Evaluation*
  - Evaluated the top-10 results of 10 sample queries used in Task 2

Task 4

The goal of this task was to use weighted NLP features to improve the results obtained in Task 3.

- *Weight Training*
  - Created a training set by manually annotating search queries
  - Used Machine Learning to learn weights of features

- *Evaluation*
  - Evaluated the top-10 results of 10 sample queries used in Tasks 2 and 3 using weights learnt during weight training

2. Implementation

This section of the report details the programmatic aspects of implementing each of the tasks outlined in Section 1.2.

2.1. Programming Tools

The following tools were used to implement the project:

- JDK v1.8
- Apache Solr
- Apache Maven
- Stanford Core NLP Library
- MIT Java Wordnet Interface Library

A more detailed list of sources and references can be found in Section 5.

2.2. Implementation Architecture

The implementation can be divided into the following two major pipelines:

1. The Information Extraction and Retrieval Pipeline
2. The Weight Learning Pipeline

The Information Extraction and Retrieval Pipeline (IERP) is a double-layered pipeline that deals with extracting information from the corpus and storing it in the index, as well tokenizing a user-supplied query and running a search against the index. Since most of the components overlap, they are both a part of the same pipeline but on separate layers.

A brief layout of the components of the IERP is given in Figure 2.1 below.

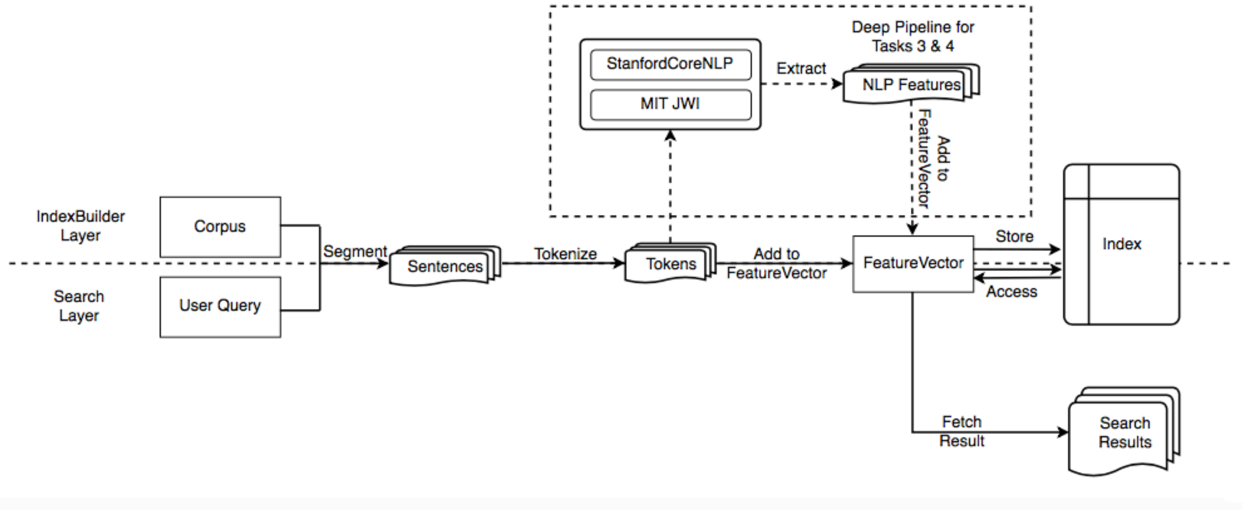

Fig 2.1 Information Extraction and Retrieval Pipeline

The Weight Learning Pipeline (WLP) describes the flow of the weight-training algorithm used to do weighted feature search in Task 4. A brief layout of its components are given in Figure 2.2. However, in order to fully understand their workings, the reader is encouraged to read Section 2.3 first.

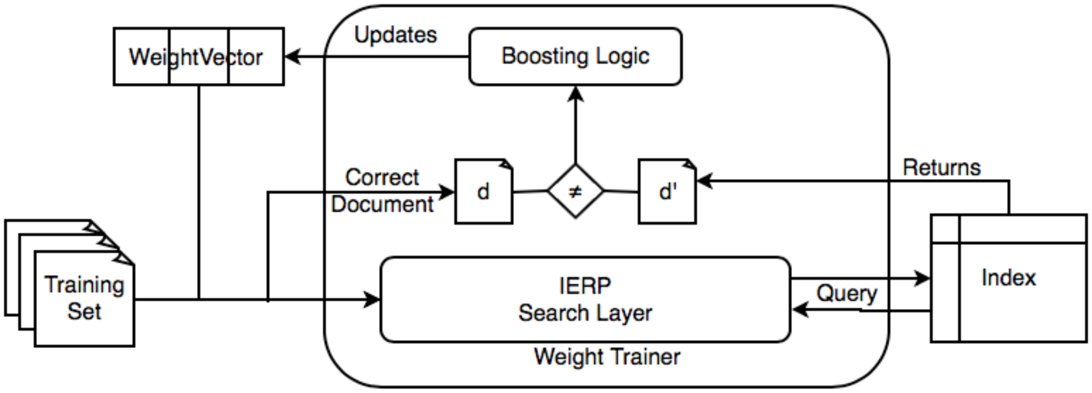

Fig 2.2 Weight Learning Pipeline

2.3. Weighted Feature Search

The primary aim of Task 4 was to add an additional layer of processing in order to refine the search results returned by the IERP pipeline for Task 3. It was decided to use weighted features instead of unweighted ones. The problem that remained was to decide on what weights needed to be used. In order to solve this problem, a weight-learning algorithm inspired by machine learning principles was written.

Weight-Learn

Input: Training set containing queries with best possible matches, max_iterations, learning_rate

Output: Weight vector for all features

```
weights <- vector with all 1’s
Until convergence or max_iterations:
    For x in training set:
        d <- x.d // correct document for x.query
        d’ <- run x.query against index
        if d.id == d’.id:
            continue
        score1 <- similarity scores for each feature of x.query against d
        score2 <- similarity scores for each feature of x.query against d’
        score <- (score1 - score2) * learning_rate
        weights <- weights + score
```

The boosting is inspired by the gradient descent algorithm. For a given query that yields a different document id from the one given in the training set, the piecewise similarity scores for every individual feature (such as words, lemmas, POS tags, etc.) are calculated. The scores against the current document are then subtracted from the scores against the correct document. The motivation behind this operation is to give more weightage to parts of the query that are more similar to the correct document than the current document. The final boosting scores are then multiplied by the learning rate and are added to the weights.

3. Analysis

This section contains the results obtained by parsing sample queries through the system. It also touches upon the problems encountered and any future improvements that could be made.

3.1. Results and Error Analysis

The following queries were run through the system:

1. A drop in gasoline prices obviously helps with consumer finances
2. Bananas remained one of the chain's top selling products because only high-quality specimens
3. Queensland farmers say they fear for their livelihoods if the state's controversial water reforms go ahead
4. Farming is a tough unforgiving life.
5. The price is taking a toll on farmers.
6. Grazing continues as the highest value sector in farm production.
7. It is expected to increase Australias GDP by $16 billion.
8. Some of the decline in gas prices is due to weakness and worry about the economy.
9. Gas prices spiked as oil refineries were forced offline.
10. For a farmer rain cold or suns do not matter much.

A significant improvement was observed when running them against the index produced in Task 3 than Task 2. Further improvement was also observed when run using weighted features in Task 4. We will examine and analyze three queries out of the list of ten below.

Query 1: For a farmer rain cold or suns do not matter much.

Task 2

Id	Score	Sentence
132_4	14.48	What wed really like to be able to do is to tell the farmer whether the climate conditions in the coming six months would be really good for growing or not Mr McIntosh said
2218_3	13.12	We believe that any solution that leaves a monopoly in place in Australia Canada wherever it is no matter how you shuffle the boxes around or shift responsibility from one organisation to another a monopoly is not a permanent solution spokesman Steve Mercer said
2133_4	12.14	If the Government is fair dinkum about helping us well the first thing they can do is pay for any water that were not going to get you know for a farmer to be lumbered with a $20000 water bill for example and not get any production out of it it makes it pretty darn hard Mr Grinter said
891_4	12.00	Buckleboo farmer Jim Hamlin says the mice are just another problem when they are already facing a cold dry start to the season

Id	Score	Sentence
2281_2	11.76	For most of her life Mrs Madden has been a dairy farmer with her late brother Len
747_5	11.49	They had good reserves from the tree and having that cold snap follow on from that rain seems to have triggered a lot of trees that were ready to go off
559_2	11.28	Some have not had a decent drop of rain since July last year forcing them to either sell or agist most of their cattle interstate
1659_1	11.25	For the first time in history cotton growers have been offered a full refund on their seed costs if they are forced to plough in their crop due to a lack of rain
806_12	10.99	They are very concerned about some of Australias pest and diseases in particular fruit fly but we've done a lot of work on fruit fly in the sense if we do cold disinfestation treatment it will kill all fruit fly he said
1467_5	10.99	We've been getting a reasonable volume of fruit out of the New South Wales growing areas and as they've gone through a very cold winter their supplies now do dwindle away he said

Task 3

Id	Score	Sentence
1839_3	168.38	Brian Clifford from the Cooma Rural Lands Protection Board says governments can not do any more now for farmers unless they can make it rain
1341_1	158.11	Farmers in Victoria Wimmera Mallee grain growing region say if it does not rain within two weeks their crops will fail
375_7	156.25	But perhaps the real significance is what we can't control and that is the rainfall out there and that is really going to be the driving factor of whether farmers have a good year or not
1513_3	155.81	After floods swept through the region in January and a very late start local farmer Trevor de Landgrafft says farmers are now relying heavily on rain in coming months
120_2	154.44	One farmer says his dams have filled for the first time in a decade with more than a hundred millimetres of rain falling
858_3	154.25	Grain and sheep farmer Bernard Gross from Drung Drung says if winter rainfall does not return to normal the future will be bleak
1626_4	154.15	Despite a few millimetres of rain overnight Ouyen farmer Alan Crook says there is now almost nothing left in his paddock worth harvesting
1502_3	153.46	Farmers say this rain will not break the drought but will keep their crops alive for another couple of weeks
1362_1	152.68	West Australian farmers are hoping for rain this week and also have some financial relief in sight with the State Government unveiling a $ 5 million drought package
1540_4	152.32	Mukinbudin farmer Chris Geraghty says the rain has been enough to bring some crops back to life and will improve pastures

Task 4

Id	Score	Sentence
1839_3	1109.81	Brian Clifford from the Cooma Rural Lands Protection Board says governments can not do any more now for farmers unless they can make it rain
1341_1	1071.10	Farmers in Victoria Wimmera Mallee grain growing region say if it does not rain within two weeks their crops will fail
858_3	1049.75	Grain and sheep farmer Bernard Gross from Drung Drung says if winter rainfall does not return to normal the future will be bleak
1513_3	1041.40	After floods swept through the region in January and a very late start local farmer Trevor de Landgrafft says farmers are now relying heavily on rain in coming months
120_2	1036.17	One farmer says his dams have filled for the first time in a decade with more than a hundred millimetres of rain falling
375_7	1035.37	But perhaps the real significance is what we can't control and that is the rainfall out there and that is really going to be the driving factor of whether farmers have a good year or not
1502_3	1033.74	Farmers say this rain will not break the drought but will keep their crops alive for another couple of weeks
1626_4	1032.91	Despite a few millimetres of rain overnight Ouyen farmer Alan Crook says there is now almost nothing left in his paddock worth harvesting
1540_4	1023.86	Mukinbudin farmer Chris Geraghty says the rain has been enough to bring some crops back to life and will improve pastures
1362_1	1021.17	West Australian farmers are hoping for rain this week and also have some financial relief in sight with the State Government unveiling a $ 5 million drought package

Query 2: Gas prices spiked as oil refineries were forced offline.

Task 2

Id	Score	Sentence
1367_3	13.30	Soaring oil prices have forced up the cost of synthetics leading to more demand for wool
276_1	10.83	Global oil prices continue to climb to record heights as Australias grain growers prepare for this years planting season
300_11	10.83	Its fair to say that as oil prices increase the interest in ethanol plants has increased markedly he said

Id	Score	Sentence
451_3	10.83	His concern comes as a world expert on energy markets Claude Mandil warns that with strong demand and falling supplies oil prices could eventually rise above $US100 a barrel
476_9	10.83	He’s about the only person in Australia who doesn’t see increasing oil prices as a major issue she said
1791_1	10.83	The NRMA is calling on the Australian Competition and Consumer Commission to investigate what it alleges is price gouging by Australias oil companies as fuel prices fall
140_3	8.62	Greg Rodert who grows potatoes for the fresh market in Bordertown says the below cost of production prices have forced many growers to cut back on the amount they produce
187_3	10.83	John Howie from the Australian Jewellers Association says at current gold prices manufacturers will be forced to up the price of jewellery and get their designs made overseas at a lower cost
643_3	10.83	Managing director Dane Hudson says the company was forced to write down stocks because of the recent decline in bulk wine prices
715_3	10.83	Vince Phillips from South East Fibre Exports at Eden in New South Wales says world prices have been forced 20 per cent below Australian,production costs

Task 3

Id	Score	Sentence
300_11	407.44	It fair to say that as oil prices increase the interest in ethanol plants has increased markedly he said
476_9	406.75	He is about the only person in Australia who does’nt see increasing oil prices as a major issue she said
1791_1	404.44	The NRMA is calling on the Australian Competition and Consumer Commission to investigate what it alleges is price gouging by Australia oil companies as fuel prices fall
300_6	404.04	The enormous legacy Margaret Thatcher has left to Britain she given the price signals that oil is going to be more expensive then they are not nearly so dependent on cheap fuel as we are
1317_1	399.18	The price of the woven polybags used to package seed wool and fertiliser is set to rise due to soaring crude oil prices
1748_5	399.18	Also what the value of the oil coming out and canola oil is reasonably highly priced at the moment particularly with the underlying biodiesel demand - then you look at the demand and the price that might be achieved for the meal and at the moment there certainly going to be domestic outlets for canola meal in Australia with the current feed situation out of this drought
476_7	398.49	Meanwhile Treasurer Peter Costello told a media conference yesterday the only way fuel prices will come down is if the cost of crude oil falls

Id	Score	Sentence
28_9	398.41	Record soy bean harvests in the US Ukraine and Brazil mean there is a large global supply of the canola alternative which will put pressure on oil seed prices in Australia
1077_2	398.41	Analysts say it is only a matter of time before crude oil hits $ US100 a barrel leading to bowser prices of about $ 2 a litre
1253_5	398.41	But everybody seems to be talking about alternative fuels with crude oil prices hitting new highs this week

Task 4

Id	Score	Sentence
300_11	2872.21	It fair to say that as oil prices increase the interest in ethanol plants has increased markedly he said
476_9	2868.67	He is about the only person in Australia who does'nt see increasing oil prices as a major issue she said
1791_1	2856.85	The NRMA is calling on the Australian Competition and Consumer Commission to investigate what it alleges is price gouging by Australia oil companies as fuel prices fall
300_6	2854.55	The enormous legacy Margaret Thatcher has left to Britain she given the price signals that oil is going to be more expensive then they are not nearly so dependent on cheap fuel as we are
28_9	2831.08	Record soy bean harvests in the US Ukraine and Brazil mean there is a large global supply of the canola alternative which will put pressure on oil seed prices in Australia
1077_2	2831.08	Analysts say it is only a matter of time before crude oil hits $ US100 a barrel leading to bowser prices of about $ 2 a litre
1253_5	2831.08	But everybody seems to be talking about alternative fuels with crude oil prices hitting new highs this week
1317_1	2829.76	The price of the woven polybags used to package seed wool and fertiliser is set to rise due to soaring crude oil prices
476_7	2826.22	Meanwhile Treasurer Peter Costello told a media conference yesterday the only way fuel prices will come down is if the cost of crude oil falls
1748_5	2818.56	Also what the value of the oil coming out and canola oil is reasonably highly priced at the moment particularly with the underlying biodiesel demand - then you look at the demand and the price that might be achieved for the meal and at the moment there certainly going to be domestic outlets for canola meal in Australia with the current feed situation out of this drought

In both these cases, we see significant improvement of results at every iteration of the Task pipeline. The "score" field is assigned by Solr that internally uses Lucene's scoring formula which is, at its core, cosine similarity.

3.2. Problems Encountered

The following problems were encountered:

1. The WordNet features (such as hypernyms, hyponyms, etc.) that were being generated were more often than not, being generated for the wrong synset. In order to remedy this, we implemented the Lesk algorithm to give the best sense for a given word and noticed significant improvement in the results.

2. Initially, the weight-training algorithm in Task 4 was not converging due to very minor differences in weights of a very small order of magnitude. This was fixed by adding a parameter for the maximum number of iterations.

3.3. Pending Issues

While significant improvement was observed after implementing Tasks 2 and 3, the only pending issues left with the project relate to potential improvements that could not be implemented during this version due to time constraints:

1. Currently, individual words are all weighed the same. However, we feel that this doesn't model the real world well and can be improved upon because certain words are typically more important than others.

2. The system currently does only syntax-based matching. We believe that incorporating semantic features such as thematic roles will also significantly boost the quality of search for ambiguous queries (although it has been attempted to remedy a part of this issue by incorporating word-sense disambiguation using the Lesk Algorithm).

3.4. Potential Improvements

The two major potential improvements based on the functionality that were not implemented due to time constraints (as outlined in Section 3.4) are as follows:

1. In order to deal with individual weighing of words, we wanted to create a weight tagger that would assign weights to individual tokens. The assignment of weights would be on the basis of the parse tree and POS tags. For instance, in the following sentence: "largest pandas in the world", the noun-phrase "largest pandas" needs to be weighed more because we would prefer results relating to large pandas in general as opposed to other animals "in the world". Similarly, within the noun-phrase "largest pandas", the word "pandas" needs to be weighed more because an article about medium-sized "pandas" are more preferable to articles about the "largest" anteater. The weights can be learnt by analyzing annotated parse trees of sample queries and therefore requires some significant manual effort.
2. A system that does searching based on only syntactic features will have its limitations with regards to queries that do not explicitly contain what the user is looking for within the

keywords. For instance, the system will not be able to understand queries such as "all buildings that are taller than fifteen meters". In order to understand such types of queries, a semantic parse engine needs to be built that can understand the meaning behind what the user is asking.

Another potential improvement could be incorporating the user's search history in order to further refine what results to show (just like Google). A set of beliefs could be maintained regarding what the user would be more likely to be search for and the set of search results could be filtered according to those beliefs.

4. Conclusion

In summary, the following findings were concluded by us during this project:

- Usage of NLP features greatly boost the quality of search results
- Weighted feature boosting may or may not yield better results, depending on the data
- In the future, it would be a better idea to incorporate weighting based on parse structures as well as semantic features for further improvement in search.

5. References

1. Australian Broadcasting Commission 2006 Corpus (http://www.nltk.org/nltk_data/)

2. Apache Solr (http://lucene.apache.org/solr/)
3. Apache SolrJ (https://wiki.apache.org/solr/Solrj)
4. Stanford Core NLP (https://stanfordnlp.github.io/CoreNLP/)
5. MIT Java Wordnet Interface (https://projects.csail.mit.edu/jwi/)